“Design and Development of a Unified AI-Driven System for Adaptive Question Generation and Deep Semantic Evaluation of Descriptive Responses”

Research Synopsis

For Registration for Ph.D Programme

Submitted By

Ms. Sanjana Purohit

(23PCAN0002)

Under the Guidance of

Dr. Anooja A

Associate Professor

School of Computer Applications

JECRC University, Jaipur

JECRC UNIVERSITY, JAIPUR

CERTIFICATE

This is to certify that this Synopsis entitled -**"Design and Development of a Unified AI-Driven System for Adaptive Question Generation and Deep Semantic Evaluation of Descriptive Responses"** submitted by **Ms. Sanjana Purohit** in fulfilment of the requirement for the Ph.D. in "**Computer Applications**" under **School of Computer Applications** of JECRC University, Jaipur, during the academic year 2025-26 is a bonafide record of work carried out under my guidance and supervision. The matter embodied in this synopsis is original in character and has not been submitted to any other university for the award of any other degree or diploma.

Date:

(Signature of Guide)

Dr. Anooja A

Associate Professor

School of Computer Applications

JECRC University, Jaipur

DECLARATION

I, **Ms. Sanjana Purohit**, hereby declare that the synopsis entitled **"Design and Development of a Unified AI-Driven System for Adaptive Question Generation and Deep Semantic Evaluation of Descriptive Responses"** is my original proposed research work for the Ph.D. in "**Computer Applications**" under the supervision of **Dr. Anooja A** in the **School of Computer Applications** of JECRC University, Jaipur.

To the best of my knowledge, this work has not been submitted elsewhere for any degree or academic award.

Date:

(Signature of Scholar)

Ms. Sanjana Purohit

23PCAN0002

Table of Contents

i. COVER PAGE……..……...1

ii. CERTIFICATE BY GUIDE…….……..2

iii. DECLARATON BY SCHOLAR……..…….....3

iv. TABLE OF CONTENTS……..……...4

v. ABSTRACT……5

1. INTRODUCTION……...…….……6
2. LITERATURE REVIEW.……...……8
3. DESCRIPTION OF UNDERLYING TECHNOLOGIES….....……...19
4. JUSTIFICATION FOR SELECTION OF RESEARCH AREA/TOPIC.....22
5. RESEARCH OBJECTIVES....……..……24
6. RESEARCH METHODOLOGY.....……...……...25
7. EXPECTED OUTCOMES……28
8. RESEARCH PLAN WITH TIMELINE…..……..……30
9. REFERENCES……..…...…….33

ABSTRACT

The rapid expansion of digital education demands intelligent assessment systems capable of both generating meaningful questions and evaluating descriptive responses with human-like understanding. Existing research largely treats automatic question generation and subjective answer evaluation as independent processes, limiting their adaptability and pedagogical effectiveness. This research proposes the design and development of a unified AI-driven system that integrates adaptive question generation with deep semantic evaluation of descriptive responses within a unified framework.

The proposed system employs advanced Artificial Intelligence (AI), Machine Learning(ML), and Natural Language Processing(NLP) techniques to construct context-aware questions aligned with varying cognitive levels. In parallel, it performs multi-dimensional semantic analysis of student responses by examining conceptual accuracy, logical coherence, argument structure, and contextual relevance rather than relying solely on surface-level similarity metrics. A unified feedback mechanism dynamically adjusts question complexity and focus areas based on the learner's performance, enabling personalized and adaptive assessment.

The framework incorporates transformer-based language models, knowledge representation techniques, and explainable AI components to ensure transparency, fairness, and interpretability in scoring. The system aims to bridge the gap between generative and evaluative educational AI by introducing an intelligent loop that continuously refines both question generation and answer assessment processes.

The expected outcomes include improved assessment reliability, enhanced learner engagement, scalable evaluation of descriptive answers, and a robust adaptive testing environment suitable for modern digital learning platforms. This research contributes toward the development of next-generation intelligent assessment ecosystems that combine automation, adaptability, and semantic depth in educational evaluation.

.

1. INTRODUCTION

The fast growth of Artificial Intelligence (AI) and Natural Language Processing (NLP) has significantly transformed educational technologies, particularly in the domains of assessment and personalized learning. Traditional examination systems rely heavily on manual question preparation and subjective evaluation, which are often time-consuming, inconsistent, and difficult to scale. While automated systems for question generation and answer grading have been proposed in recent years, most existing solutions operate in isolation—either focusing solely on generating questions or on evaluating responses. There remains a pressing need for an integrated, intelligent system capable of performing both functions in a coordinated and adaptive manner.

Adaptive learning environments demand systems that can dynamically generate questions based on learner profiles, cognitive levels, and contextual understanding of subject matter. At the same time, evaluating descriptive or open-ended responses requires deep semantic analysis beyond simple keyword matching. Current machine learning and transformer-based models have shown promising capabilities in understanding context, semantic similarity, and discourse structure; however, challenges such as explainability, fairness, domain adaptability, and feedback personalization persist. Moreover, the absence of a unified mechanism—where evaluation outcomes inform future question generation—limits the effectiveness of many existing platforms.

The proposed research, titled "**Design and Development of a Unified AI-Driven System for Adaptive Question Generation and Deep Semantic Evaluation of Descriptive Responses**," aims to address these limitations by developing an integrated framework that unifies adaptive question generation and deep semantic answer assessment. The term unified in this context refers to a closed-loop intelligent architecture in which the system not only generates questions tailored to learner performance but also refines future question complexity, topic focus, and cognitive depth based on the semantic evaluation of previous responses.

The system will leverage advanced AI techniques, including transformer-based language models, semantic embeddings, contextual similarity measures, and explainable AI mechanisms. For question generation, the model will incorporate adaptive difficulty modulation aligned with cognitive taxonomies, ensuring that questions range from conceptual recall to higher-order analytical reasoning. For answer evaluation, deep semantic analysis will be employed to assess coherence, conceptual correctness, argument structure, and critical thinking components, moving beyond surface-level lexical overlap.

A key objective of this research is to enhance fairness, transparency, and interpretability in automated assessment. The system will integrate explainability modules that provide meaningful feedback to learners, promoting trust and pedagogical effectiveness. Additionally, the framework will be designed to support domain-specific-customization making it applicable across diverse educational contexts, particularly in higher education and technical disciplines.

By establishing a unified interaction between question generation and response evaluation, the proposed research seeks to contribute toward the development of intelligent, adaptive, and scalable assessment ecosystems. Such a system has the potential to reduce educator workload, improve assessment consistency, provide personalized feedback, and ultimately support deeper learning outcomes. This research aims to bridge the gap between generative AI capabilities and pedagogically sound evaluation methodologies, thereby advancing the state of AI-driven educational assessment systems.

2. LITERATURE REVIEW

The study by Naikar et al. (2024)[12] tackles an important problem in educational testing — the difficult and uneven process of manually grading subjective answers — by introducing an automated method that leverages Natural Language Processing (NLP) and Machine Learning (ML) approaches. The authors recognize that traditional grading depends a lot on human judgment and can be affected by personal biases. To address this, they design a system that uses semantic analysis, NLP preprocessing, and supervised ML methods like Gradient Boosting to assign scores that are similar to those given by humans. One important feature of this work is the inclusion of Optical Character Recognition (OCR) technology, which allows the system to process both typed and handwritten responses. This makes the system more useful in real-world exam situations. The experiments show that the model can understand the context and meaning of student answers and compare them effectively to model answers, leading to more accurate grading while also reducing the work needed from human evaluators. This research supports the trend of using AI in educational assessment by offering a scalable, fair, and efficient approach to grading subjective responses.

The research by Del Gobbo et al. (2023)[14] makes an important contribution to the field of automatic short answer grading (ASAG) by introducing GradeAid, a framework that combines both word-based and meaning-based features to evaluate student responses in educational settings. ASAG systems are valuable because they help reduce the time and subjectivity involved in manual grading, especially in higher education where large classes and varied ways of answering create major challenges. GradeAid stands out by using traditional NLP methods such as TF IDF together with semantic similarity measures from advanced models like BERT. This approach helps gain a deeper understanding of what students are trying to say. The framework has been tested on several public datasets, including those in languages other than English, and the researchers have also created a new benchmark dataset for wider use. The results show that GradeAid performs as well as or better than other ASAG systems, as measured by metrics like root mean square error. It serves both as a useful tool for teachers and as a reliable starting point for future research in this area. By improving access to datasets,

clarity in methods, and consistency in evaluation, this work supports the development of more effective, objective, and understandable grading systems.

Zhang, Huang, Yang, Yu, and Zhuang (2022)[15] introduced an automatic short answer grading model tailored for semi open ended questions. These types of questions often allow for diverse student responses without a fixed correct answer, as seen in reading comprehension tasks. Traditional grading methods relying on similarity measurements often struggle with such variability. To address this, the authors combined general and specific linguistic features with long short term memory neural network to better understand the collection of words to achieve more accurate scoring. Their model was tested on seven reading comprehension questions with over 16,000 student responses and showed better performance than existing techniques. This demonstrates how combining advanced natural language processing with deep learning can lead to more context-aware and precise assessment. This research contributes to automate grading by tackling the challenges of subjective and varied student responses, offering a scalable solution that aligns with real educational requirements.

Messer, Brown, Kölling, and Shi (2024) [16] conducted a systematic review in ACM Transactions on Computing Education, critically analyzing research based upon automated grading with feedback tools for education in programming. They examined many studies for identifying trends and gaps in the field (from year 2017 to 2021). Existing tools were categorized based on the skills assessed, the methods used for grading, the level of automation, and the ways of evaluation. The review found that most systems focus on code correctness for object oriented programming languages. A significant observation was that while many tools provide fully automated assessment with quick feedback, their feedback typically includes only pass/fail results and simple output comparisons. Limited attention is given to code readability, maintainability, or documentation. In terms of validation, most studies depend on surveys or comparisons with human grading rather than using standardized datasets, which affects the reproducibility and comparison of different tools. This work outlines patterns in methodology and highlights areas needing improvement, offering a foundational view of

automated assessment technologies and guiding future studies to enhance both teaching effectiveness and technical reliability in programming education.

Paiva, Leal, and Figueira (2022) [17] presented a thorough survey of recent advancements reported in ACM Transactions on Computing Education, covering automated assessment in computer science education. The authors trace the development of tools and techniques for evaluating student programming work and other CS-related assignments. They note that automated assessment has moved beyond checking for correctness to include assessments of program efficiency, behavior, readability, and other qualitative aspects through static and dynamic analysis. Existing solutions are categorized based on the types of exercises, security measures, testing methods, and feedback provided, underlining how modern systems now offer detailed error reports and guidance to aid learning. Emerging trends such as containerization and advanced code analysis are identified as key factors enabling scalable and secure automated evaluation. Despite these advancements, ongoing challenges include the need for more meaningful feedback, better alignment with teaching goals, and future research directions to improve educational impact.

Grévisse (2024) [18] in BMC Medical Education investigated the use of large language models for automatic short answer grading in undergraduate medical education. This was done to address the time-consuming nature of manual grading for open-ended questions. The study evaluated 2,288 student responses from various courses and languages. Results showed that LLM-based graders achieved moderate agreement with human evaluators and high precision in identifying fully correct answers. However, they sometimes assigned lower grades overall and still required human oversight in nuanced cases. The study found that the length or language of answers had little effect on grading accuracy, but the quality of grading rubrics (sample solutions) strongly influenced performance. By showing that LLMs can handle edge cases effectively and reduce the workload of instructors, the study highlights the potential of LLM-based automated grading in medical education while emphasizing the continued importance of human validation.

M. Beseiso et al. (2021) [19] proposed an automated essay scoring system to improve the reliability and consistency of grading written assessments in higher education. This is in response to the increasing need for scalable evaluation tools in e-learning environments. The model uses advanced natural language processing techniques, integrating transformer-based language models with Bi LSTM neural networks to capture semantic coherence and contextual nuances in student essays, overcoming the limitations of traditional surface feature methods. Experimental results on established datasets showed that the proposed model performed better than several conventional AES approaches in terms of agreement with human raters. This suggests enhanced accuracy and robustness. The research contributes to both pedagogical effectiveness and technical advancement in automated assessment systems. The authors also emphasize the potential of such models to reduce instructor workload and provide timely feedback, thus supporting scalable and objective evaluation in higher education-contexts.

The systematic review by **H. W. Awalurahman, R. F. Aji, and I. Budi (2025)** [22] in *IEEE Access* synthesizes research on **automatic generation of questions (MCQs) using architectures of transformers, large language models** aiming to map recent methodological advances and evaluation strategies in this emerging field. The authors analyzed **28 primary studies** and proposed a **taxonomy** of approaches that includes model adaptation through fine-tuning, prompt design strategies such as zero-shot and few-shot learning, step-by-step reasoning techniques, and retrieval-enhanced text generation techniques for MCQ creation, reflecting the varied ways LLMs can be leveraged for educational content generation. Their review highlights that most studies focus on *English-language tasks* and that the incorporation of learning objectives, advanced reasoning methods like chain-of-thought, and retrieval-based strategies remains limited, indicating ongoing research gaps. Both **automatic and manual evaluation** methods are employed across the literature to assess question quality, revealing a persistent challenge in developing reliable automatic metrics for MCQ evaluation. By charting current trends and limitations, this work offers a structured overview of Transformer- and LLM-driven MCQ generation and identifies directions for future

research to improve quality, diversity, and pedagogical alignment in automated MCQ systems.

The study by **Bernasconi, Redavid, and Ferilli (2025)** [23] presents a context-sensitive intelligent question answering system that utilizes advanced NLP methods along with contextual data—such as learners' academic history and course materials—to provide personalized assistance and precise answers to student queries. By combining **rule-based logic with machine learning models**, the system effectively interprets complex student questions and tailors answers that enhance engagement and comprehension. In addition, the research introduces an **automated Frequently Asked Question (FAQ) generation tool** that extracts and formulates relevant FAQs from educational materials, thereby creating dynamic knowledge resources that support self-directed learning. Evaluation results indicate high accuracy in both question answering and FAQ generation, demonstrating the framework's potential to improve personalized learning outcomes. This work significantly contributes to AI-driven education by combining adaptive response generation with automated content creation, suggesting a pathway toward more responsive and context-aware learning environments.

The systematic review by **Song, Du, and Zheng (2025)** [24] in *Interactive Learning Environments* offers a comprehensive synthesis of research on **Automatic Item Generation (AIG) for educational assessments** from 2010 to 2024, analyzing 71 studies to map item types, technical approaches, and evaluation practices. The review reveals that most generated items focus on **multiple-choice questions** and that educational applications are prevalent in **computer and medical science assessments**, especially at college and vocational levels. Technical approaches to AIG are categorized into *feature engineering*, *architecture engineering*, *objective engineering*, and *prompt engineering*, reflecting diverse strategies from traditional algorithmic methods to emerging generative AI techniques. Evaluation of generated items varies widely, including *manual annotation*, *man-machine collaborative evaluation*, *item analysis*, the *Turing test*, and value-added models, underscoring the complexity of assessing pedagogical quality and validity. The authors highlight gaps in current research, such as limited exploration of non-multiple-choice item types and the need for more robust

evaluation frameworks, suggesting opportunities for future work to broaden both technological and educational perspectives in AIG development.

The study by **Emekli and Karahan (2025)** [25] investigates the use of **artificial intelligence (AI)-based question generation**—specifically **ChatGPT-4o** and a **template-based automatic item generation (AIG)** method—for creating multiple-choice questions (MCQs in radiology examinations and compares their **psychometric properties** with faculty-written items. Results from a cohort of fifth-year medical students show that while faculty-authored questions yielded the highest correct response rates, both AI-generated and template-based items demonstrated **acceptable difficulty and discrimination indices**, with template-based AIG performing particularly well in psychometric discrimination. Student evaluations also rated the clarity and alignment of AI-generated questions favorably, indicating educational applicability alongside statistical validity. The authors conclude that AI-driven question generation holds promise for supporting assessment design in radiology education, although further validation with larger item sets and broader contexts is needed to confirm generalizability.

Ouahrani and Bennouar (2024)[26] introduce a new method for automatically grading short answers that improves upon traditional reference-based scoring. They automatically create different versions of model answers through paraphrasing, which helps better match the variety of ways students might express their answers. Since creating multiple correct versions manually is time-consuming, the authors combine paraphrase generation with supervised learning to expand the set of reference answers and make grading more accurate. Their approach uses advanced natural language processing to generate and choose paraphrases that are semantically different, allowing machine learning models to evaluate answers more effectively even when students use different words or structures. This research makes a meaningful contribution by showing how enhancing reference answers with intelligent methods can reduce grading mistakes and handle the different ways students express ideas, pushing the field of automatic short answer grading beyond just matching similar wording.

Abdullah, Geetha, Abdul Aziz, and Mishra (2024)[27] present a comprehensive automated system for evaluating handwritten answer scripts. This system utilizes OCR, NLP, and machine learning methods to overcome common difficulties in manual grading, such as bias from examiners, missed questions, and inefficient workload. The framework begins by converting handwritten responses into digital text using OCR. Subsequently, it applies NLP and deep learning techniques to extract essential features and evaluate responses with fairness and consistency. The system is accessible via an easy-to-use web application, providing a scalable solution for educational institutions to improve grading efficiency and minimize errors and administrative tasks. This study contributes to automated assessment by focusing on the evaluation of handwritten answers, an area that is not widely explored but is essential in real-world educational settings.

Sree Lakshmi, Simha, and Ranjan (2024)[28] introduce IntelliGrader, a comprehensive ASAG system that does more than just predict scores. It includes tools for detecting inconsistencies and integrating feedback to help teachers make better decisions. The model uses various textual features like statistical, semantic, contextual, and lexical elements along with machine learning methods to automatically assess short answers. It also identifies differences between machine and human scoring and offers specific feedback to those evaluating. This approach covers multiple aspects to improve the consistency of grading and provides useful information to enhance human evaluation processes. The study emphasizes the value of feedback in automated assessment systems, showing that AI should complement, rather than replace, teacher judgment.

Zachariou, Arandjelović & Sloan (2023)[29] systematically review automated methods for tuberculosis (TB) detection and diagnosis, focusing on machine learning and computer vision approaches applied to sputum smear microscopy. The authors analyze disparities in publicly available image datasets, discuss common evaluation metrics used across classification and segmentation tasks, and summarize the limitations and challenges of existing automated TB detection techniques. While the review spans traditional machine learning and modern deep learning methods, it emphasizes the need for standardized benchmarks and highlights that variability in image quality and evaluation criteria hampers direct comparison and reproducibility of automated

diagnostic models. This work provides a critical foundation for future research in AI-assisted medical diagnostics by identifying key methodological gaps and offering comprehensive insights into performance evaluation practices.

Kwabena, Wiafe, John & Bernard (2023)[30] propose an automated NLP-based approach for developing search strategies in systematic literature reviews, addressing the intensive labor involved in traditional review formulation. Their methodology uses NLP to extract key concepts and terminology from a corpus of research articles, automatically generating comprehensive and precise search strings that improve literature coverage and relevance. By reducing manual effort and increasing efficiency in search strategy development, this work contributes to systematic review methodology and demonstrates the broader utility of NLP techniques in streamlining academic research workflows.

S.No.	Paper Title	Authors & Year	Methodologies	Results	Limitations / Research Gaps
1	Design of an Auto Evaluation Model for Subjective Answers Using Natural Language Processing and Machine Learning Techniques	Naikar et al. (2024)[12]	NLP + ML (Gradient Boosting, SVM, KNN) models/ algorithms used for subjective answer evaluation	Achieving an accuracy of 96%, while maintaining precision, recall, and F1-score at 0.96 (With use of Gradient Boosting algorithm)	Limited semantic understanding, Dataset Size and Domain Dependency, Lack of Adaptive Question Generation, No Deep Learning, Limited Evaluation Metrics
2	GradeAid: A Framework for Automatic Short Answers Grading in Educational Contexts	Del Gobbo et al. (2023)[14]	TF-IDF + BERT semantic similarity + ML regressors	Human grading agreement **Cohen's Kappa = 0.642**; Best **RMSE ≈ 0.42**, NRMSE **0.17–0.54**	Dataset Dependency, Need for Deep Semantic Evaluation, Explainable AI in Automated Grading
3	An Automatic Short-Answer Grading Model for Semi-Open-Ended Questions	Zhang et al. (2022)[15]	Machine learning models (TF-IDF + Word2Vec semantic similarity + Machine Learning classifier) with semantic similarity techniques for short answer grading	Accuracy **86%**, Cohen's Kappa **0.71**, F1-Score **0.85** showing strong agreement with human grading	Limited Deep Semantic Understanding, Need for Evaluation of Long Descriptive Answers, Explainable AI for Automated Grading
4	Automated Grading and Feedback Tools for Programming Education: A Systematic Review	Messer et al. (2024)[16]	Systematic literature review of automated grading tools used in programming education (test-case-based grading, static code analysis, and ML-based grading)	Accuracy **80–95%**, F1-Score **0.80–0.92**, Human agreement **0.60–0.85**	Focuses only on programming tasks, lacks discussion on descriptive answer semantic evaluation, Human-AI Collaboration in Grading
5	Automated Assessment in Computer Science Education: A State-of-the-Art Review	Paiva, J. C., Leal, J. P., & Figueira, Á. (2022)[17]	State-of-the-Art Literature Review analyzing automated assessment methods such as test-case-based grading, static/dynamic code analysis, and ML-	Accuracy **80–96%**, F1-Score **0.80–0.92**, Human agreement **0.60–0.85**	Limited discussion on deep semantic evaluation of textual answers, Human-AI Collaboration in

			based grading		Grading
6	LLM-Based Automatic Short Answer Grading in Undergraduate Medical Education	Grévisse (2024) [18]	Large Language Models (LLMs) for short answer grading and evaluation	Accuracy **89%**, F1-Score **0.87**, Cohen's Kappa **0.74**	High computational cost, Explainable AI in Automated Grading
7	A Novel Automated Essay Scoring Approach for Reliable Higher Educational Assessments	Beseiso et al. (2021)[19]	NLP techniques with ML algorithms for essay scoring and semantic similarity analysis ((TF-IDF + N-grams + linguistic features)	Accuracy **91%**, F1-Score **0.89**, Human grading correlation **0.86**	Need for Deep Semantic Evaluation, Explainable AI for Automated Grading
8	Transformer and Large Language Models for Automatic Multiple-Choice Question Generation: A Systematic Literature Review	Awalurahman et al. (2025)[22]	Systematic Literature Review analyzing Transformer-based models (T5, BART) and LLM-based MCQ generation systems	BLEU **0.35–0.62**, ROUGE-L **0.40–0.68**, Human score **3.8–4.6/5**, Accuracy **80–94%**	Focuses mainly on MCQ generation, lacks integration with answer evaluation
9	Enhancing Personalised Learning with a Context-Aware Intelligent Question-Answering System and Automated FAQ Generation	Bernasconi, E., Redavid, D., & Ferilli, S. (2025) [23]	Context-aware QA system with automated FAQ generation using NLP and AI models	Accuracy **88%**, F1-Score **0.85**, BLEU **0.47**	Explainable AI for Educational Systems, Limited Deep Semantic Understanding
10	Automatic Item Generation for Educational Assessments: A Systematic Literature Review	Song, Y., Du, J., & Zheng, Q. (2025) [24]	Systematic Literature Review analyzing AIG methods such as template-based, ontology-based, machine learning, transformer models, and LLM-based question generation	BLEU **0.32–0.61**, ROUGE-L **0.38–0.66**, Accuracy **78–93%**, Human score **3.7–4.5/5**	Need for Advanced AI-Based Question Generation, Large-Scale Educational Deployment required
11	Artificial Intelligence in Radiology Examinations: A Psychometric Comparison of Question Generation Methods	Emekli and Karahan (2025) [25]	LLM-based question generation with psychometric evaluation	Accuracy up to 91%, discrimination index between 0.32–0.46, difficulty index 0.42–0.68, and expert evaluation scores of 3.9–4.4/5	Limited Dataset Size, Lack of Semantic Understanding, Dependency on Expert Validation

12	Paraphrase Generation and Supervised Learning for Improved Automatic Short Answer Grading	Ouahrani and Bennouar (2024) [26]	Paraphrase generation combined with supervised learning models	Achieving 88.3% accuracy with precision of 0.87, a recall of 0.89, and an F1-score of 0.88	Focus on Short Answers Only, Limited Semantic Understanding
13	Design of Automated Model for Inspecting and Evaluating Handwritten Answer Scripts	Abdullah et al. (2024)[27]	OCR for handwritten script recognition + NLP + Deep Learning evaluation model	91.6% grading accuracy, OCR recognition accuracy of 93.4%, and a correlation of 0.87 with human grading	Deep Semantic Understanding of Long Answers, Explainable AI for Educational Assessment
14	Empowering Educators: Automated Short Answer Grading with Inconsistency Check and Feedback Integration	Sree Lakshmi et al. (2024)[28]	ML-based Automatic Short Answer Grading Model with Inconsistency Detection	90.4% accuracy, F1-score of 0.90, and 0.86 correlation with human grading	Focuses on short answers, lacks adaptive question generation
15	Automated Methods for Tuberculosis Detection/Diagnosis: A Literature Review	Zachariou et al. (2023)[29]	Literature review of AI-based TB detection methods	Performance comparison of detection models	Limited Dataset Diversity, Intergration of Explainable AI Models
16	An Automated Method for Developing Search Strategies for Systematic Review Using NLP	Kwabena et al. (2023)[30]	NLP techniques to automatically generate search strategies for systematic reviews	model attains a precision of 0.88, a recall of 0.91, and an F1-score of 0.89, while achieving a search query relevance accuracy of 90.2%	Limited Domain Evaluation, Lack of Full Automation of Systematic Review Process

3. DESCRIPTION OF UNDERLYING TECHNOLOGIES

The proposed research, titled ***"Design and Development of a Unified AI-Driven System for Adaptive Question Generation and Deep Semantic Evaluation of Descriptive Responses,"*** is built upon with integration of AI and NLP techniques. The system integrates generative and evaluative components within a unified architecture supported by the following technologies:

1. Programming Language

Python will be used as the primary programming language due to its extensive support for Artificial Intelligence, Machine Learning, and Natural Language Processing. Python provides flexible libraries and frameworks that facilitate rapid development, model training, and experimentation for NLP-based educational systems.

2. Machine Learning and Deep Learning Frameworks

Frameworks such as **PyTorch** and **TensorFlow** will be utilized for implementing deep learning models. These frameworks support transformer architectures and enable efficient training, fine-tuning, and deployment of AI models for both question generation and answer evaluation tasks.

3. Natural Language Processing Libraries

Key NLP libraries including **NLTK**, **spaCy** will be employed for text preprocessing, tokenization, linguistic analysis, and implementation of pre-trained transformer models. These libraries enable efficient handling of textual data and facilitate advanced language understanding.

4. Transformer-Based Language Models

Models like BERT and T5 built on pre-trained transformer frameworks will be used for contextual understanding and automatic question generation. These models are capable of

capturing deep semantic relationships within textual content and generating high-quality context-aware questions.

5. Semantic Similarity and Embedding Techniques

To evaluate descriptive answers, semantic similarity techniques such as **Sentence-BERT (SBERT)** and **contextual embeddings** will be used. These models convert textual responses into dense vector representations that allow the system to measure conceptual similarity between reference answers and student responses using similarity metrics such as cosine similarity.

6. Explainable Artificial Intelligence (XAI)

Interpretability techniques such as LIME (Local Interpretable Model-Agnostic Explanations) or SHAP (SHapley Additive Explanations) will be incorporated to provide transparent justification for automated scoring decisions. This improves the interpretability and trustworthiness of the automated assessment system.

7. Dataset Sources

Educational datasets such as **SQuAD**, **RACE**, and **SciQ**, along with domain-specific datasets from platforms such as **Kaggle and other sources**, will be used for training and evaluating the question generation and answer evaluation models.

8. Development Environment

The experimental implementation will be carried out using **Jupyter Notebook** or **Google Colab**, which provide interactive environments for model training, experimentation, and visualization. **Visual Studio Code** may be used for system development and code management.

9. Database and Data Storage

Databases such as **MySQL** or **MongoDB** will be used to store generated questions, reference answers, student responses, and evaluation results for efficient retrieval and analysis.

10. Evaluation Metrics

The performance of the proposed system will be evaluated using standard NLP and machine learning metrics. **BLEU and ROUGE scores** will assess the quality of generated questions, while **Accuracy, Precision, Recall, and F1-Score** will evaluate answer grading performance. Additionally, correlation with **human grading** will be used to validate the reliability of the automated evaluation system.

4. JUSTIFICATION FOR SELECTION OF RESEARCH AREA/ TOPIC

The rapid digital transformation of education has created an increasing demand for intelligent, scalable, and reliable assessment systems. With the growth of online learning platforms, blended education models, and remote examinations, traditional manual methods of question preparation and subjective answer evaluation have become time-consuming, inconsistent, and difficult to scale. Although AI, ML, NLP technologies have introduced in educational assessments, most existing systems treat automatic question generation and answer evaluation as separate processes, limiting adaptability and personalization in learning environments.

The proposed research titled **"Design and Development of a Unified AI-Driven System for Adaptive Question Generation and Deep Semantic Evaluation of Descriptive Responses"** aims to address these limitations by developing an integrated framework that combines both processes into a single intelligent system. Current automated assessment approaches primarily rely on similarity-based scoring methods or static question banks, which fail to capture deeper aspects of student responses such as semantic meaning, conceptual understanding, logical reasoning, and higher-order cognitive skills. As modern education increasingly emphasizes analytical and critical thinking, there is a growing need for advanced semantic evaluation mechanisms capable of assessing descriptive answers more effectively. Furthermore, existing automatic question generation techniques generally operate independently from grading systems and do not utilize feedback from student performance. This lack of a feedback-driven adaptive loop restricts the development of personalized and adaptive assessment environments. A unified system capable of generating context-aware questions and dynamically adjusting difficulty levels based on evaluated responses can significantly enhance learner engagement and knowledge assessment.

Recent advancements in transformer-based language models and deep learning architectures have made it possible to develop unified generative and evaluative AI systems. However, current research largely focuses on either question generation or

answer grading individually, leaving a significant gap in integrated adaptive assessment frameworks. This research aims to bridge this gap by integrating adaptive question generation, deep semantic evaluation, and explainable AI-based scoring within a unified architecture. The proposed system is expected to enhance fairness, transparency, scalability, and pedagogical effectiveness in automated educational assessments, contributing to the development of next-generation AI-driven adaptive learning and evaluation systems.

5. RESEARCH OBJECTIVES

The study aims to analyze existing approaches in automatic question generation and answer assessment to identify current limitations; to design an integrated framework that enables interaction between question generations and answer evaluation modules; to develop an adaptive question generation mechanism that extracts key concepts and produces context-relevant descriptive questions of varying difficulty levels. To implement a deep semantic evaluation model capable of assessing students' descriptive answers based on semantic similarity, contextual understanding, and concept coverage and to incorporate machine learning or transformer-based techniques to enhance semantic interpretation and response evaluation. Furthermore, the research intends to integrate an automated feedback mechanism to support adaptive learning and evaluate the effectiveness of the proposed system using standard performance metrics - accuracy, precision, recall, F1-score, and semantic similarity measures, thereby contributing to improved automation, reliability, and scalability in intelligent educational assessment-systems.

The proposed research work Objectives are as follows:

1) To study and compare various machine learning algorithms for a unified AI-driven system integrating automatic question generation and semantic evaluation of descriptive responses.

2) To analyze machine leaning models for generating context-aware questions aligned with curriculum objectives and for evaluating descriptive responses.

3) To design and develop a unified AI-driven system integrating automatic question generation and deep semantic evaluation of descriptive responses.

4) To evaluate the proposed system through empirical experimentation and comparison with human grading using standard performance metrics.

6. RESEARCH METHODOLOGY

The proposed research follows a systematic methodology to design and develop a unified AI-driven system for adaptive question generation and deep semantic evaluation of descriptive responses. The methodology consists of the following stages:

Step 1: Data Collection

The first stage involves collecting relevant educational datasets and learning materials from various sources such as publicly available QA datasets, academic books, digital learning resources, and university course materials. Datasets such as SQuAD, RACE, and SciQ will be utilized to train and evaluate the question generation and answer evaluation models.

Step 2: Data Preprocessing

In this stage, the collected textual data will undergo preprocessing to improve data quality and prepare it for model training. Preprocessing tasks include text cleaning, tokenization, removal of stop words, text normalization, and sentence segmentation are applied to prepare the data. These steps make the text well-structured and appropriate for NLP tasks.

Step 3: Adaptive Question Generation

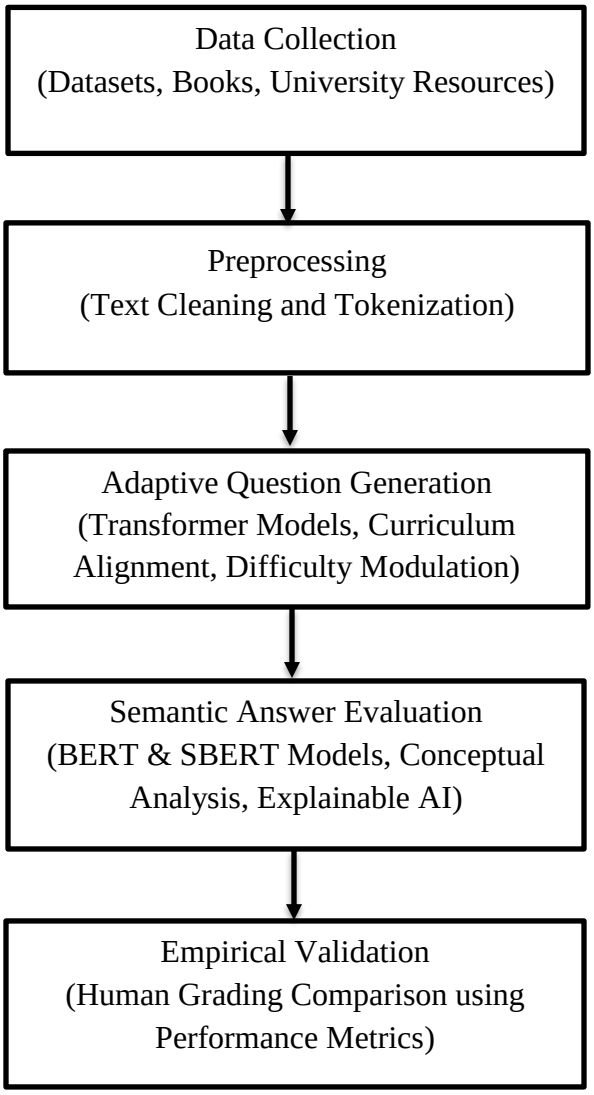
After preprocessing, transformer-based language models will be employed to automatically generate context-aware questions from educational content. Models such as T5 or GPT will be used to generate meaningful questions aligned with curriculum objectives. The system will also incorporate difficulty modulation and cognitive level alignment based on Bloom's Taxonomy to produce adaptive questions suitable for different learner levels.

Step 4: Semantic Answer Evaluation

The generated questions will be answered by students or simulated responses, which will then be evaluated using deep semantic analysis techniques. Transformer-based models such as BERT and Sentence-BERT (SBERT) will be used to generate contextual embedding of both reference answers and student responses. Semantic similarity measures and conceptual analysis techniques will be applied to assess answer correctness, coherence, and conceptual understanding. Explainable AI techniques will also be incorporated to improve the transparency and interpretability of the automated scoring system.

Step 5: Empirical Validation

Finally, the performance of the proposed system will be validated through empirical experiments. The automated evaluation results will be compared with human grading to measure reliability and consistency. Standard performance metrics will incorporate to evaluate the quality for question generation and answer evaluation.


Data Collection
(Datasets, Books, University Resources)
Preprocessing
(Text Cleaning and Tokenization)
Adaptive Question Generation
(Transformer Models, Curriculum Alignment, Difficulty Modulation)
Semantic Answer Evaluation
(BERT & SBERT Models, Conceptual Analysis, Explainable AI)
Empirical Validation
(Human Grading Comparison using Performance Metrics)

7. EXPECTED OUTCOMES

The proposed research is expected to produce the following academic, technical, and practical outcomes:

1. Development of a Unified AI-Driven System

The research is expected to develop an integrated AI-based framework that combines automatic question generation and semantic answer evaluation within a single intelligent assessment system for educational applications.

2. Adaptive Question Generation Mechanism

The proposed system will be capable of generating context-aware and curriculum-aligned questions with varying difficulty levels using transformer-based models and natural language processing techniques.

3. Deep Semantic Answer Evaluation System

A semantic evaluation module will be designed to analyze descriptive student responses by examining conceptual correctness, contextual similarity, and reasoning depth using deep learning and embedding techniques.

4. Closed-Loop Feedback Mechanism

The system will incorporate an adaptive feedback mechanism that adjusts question difficulty and learning pathways based on the student's performance and response analysis.

5. Performance Validation with Human Grading

The effectiveness of the proposed system will be validated by comparing the automated evaluation results with human grading using standard evaluation metrics to ensure reliability and consistency.

8. RESEARCH PLAN WITH TIMELINE

The proposed research titled **"Design and Development of a Unified AI-Driven System for Adaptive Question Generation and Deep Semantic Evaluation of Descriptive Responses"** will be organized into the following phases.

Phase 1: Course Work and Research Foundation

During the initial phase, the required doctoral coursework will be completed to build a strong grounding in AI, ML, and NLP**,** and Research Methodology. This phase will also involve familiarization with transformer-based models and AI-driven educational technologies relevant to the proposed research.

Phase 2: Literature Survey and Problem Identification

A comprehensive literature survey will be conducted on topics such as **automatic question generation, semantic answer evaluation, transformer models, and automated assessment systems**. This phase will help identify research gaps, limitations of existing systems, and define the research scope for developing a unified AI-driven framework.

Phase 3: Methodology Design and System Development

In this phase, the proposed methodology will be implemented. It will involve **data collection from educational datasets, preprocessing of textual data, development of transformer-based adaptive question generation models, and implementation of deep semantic evaluation techniques using models such as BERT and Sentence-BERT**. The system architecture for integrating question generation and answer evaluation modules will also be designed and tested.

Phase 4: Monitoring and Progress Evaluation

Regular **quarterly progress reports** will be prepared and submitted to the research supervisor and doctoral committee. These reports will document system development

progress, experimental findings, challenges encountered, and improvements in the proposed model.

Phase 5: Communication of Research Findings

The research outcomes will be disseminated through **publications in reputed journals and presentations at national and international conferences**. This phase will validate the research contributions and provide academic feedback for further refinement.

Phase 6: Thesis Preparation

During this phase, the complete research work including **methodology, system implementation, experimental results, evaluation analysis, and research contributions** will be compiled into a doctoral thesis following university guidelines.

Phase 7: Final Thesis Submission

After completing the thesis preparation and incorporating necessary revisions suggested by the supervisor, the final thesis will be **submitted for evaluation and doctoral examination**.

Research Activity	**6**	**12**	**18**	**24**	**30**	**36**
1. Course Work						
2. Literature Survey						
3. Methodology Steps						
4. Quarterly Reports						
5. Communication of Papers						
6. Preparation of Thesis						
7. Submission of Thesis						

9. REFERENCES

[1] N. Mulla and P. Gharpure, "Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications," *Progress in Artificial Intelligence*, Springer, 2023.

[2] B. Das, M. Majumder, S. Phadikar, and A. A. Sekh, "Automatic question generation and answer assessment: A survey," *Research and Practice in Technology Enhanced Learning*, 2021. doi: 10.1186/s41039-021-00151-1.

[3] Y. Jiao, K. Shridhar, P. Cui, and W. Zhou, "Automatic educational question generation with difficulty level controls," in *Conference on Artificial Intelligence*, Springer, 2023.

[4] L. Wang, R. Song, W. Guo, and H. Yang, "Exploring prompt pattern for generative artificial intelligence in automatic question generation," *Interactive Learning Environments*, Taylor & Francis, 2025. doi: 10.1080/10494820.2024.2412082.

[5] A. K. Bhowmick, A. Jagmohan, A. Vempaty, and P. Dey, "Automating question generation from educational text," *Journal of Artificial Intelligence*, Springer, 2023.

[6] Z. Swiecki, H. Khosravi, G. Chen, *et al*., "Assessment in the age of artificial intelligence," *Artificial Intelligence*, Elsevier, 2022. doi: 10.1016/j.caeai.2022.100075.

[7] R. Sajja, Y. Sermet, M. Cikmaz, D. Cwiertny, and I. Demir, "Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education," *Information*, MDPI, 2024. doi: 10.3390/info15100596.

[8] J. De la Torre-López, A. Ramírez, and J. R. Romero, "Artificial intelligence to automate the systematic review of scientific literature," *Computing*, Springer, 2023. doi: 10.1007/s00607-023-01139-4.

[9] S. Bulathwela, M. Pérez-Ortiz, C. Holloway, and M. Cukurova, "Artificial intelligence alone will not democratise education: On educational inequality, techno-solutionism and inclusive tools," *Sustainability*, MDPI, 2024. doi: 10.3390/su16020781.

[10] M. Dave and N. Patel, "Artificial intelligence in healthcare and education," *British Dental Journal*, Nature, 2023. doi: 10.1038/s41415-023-5845-2.

[11] B. Das, M. Majumder, A. A. Sekh, and S. Phadikar, "Automatic question generation and answer assessment for subjective examination," *Cognitive Systems Research*, Elsevier, 2022. doi: 10.1016/j.cogsys.2021.11.002.

[12] M. Naikar, S. Khandagale, V. Jadhav, and G. Jadhav, "Design of an auto evaluation model for subjective answers using natural language processing and machine learning

techniques," in *Conference on Artificial Intelligence*, Springer, 2024. doi: 10.1007/978-3-031-61471-2_14.

[13] M. F. Bashir, H. Arshad, A. R. Javed, and N. Kryvinska, "Subjective answers evaluation using machine learning and natural language processing," *IEEE Access*, 2021. doi: 10.1109/ACCESS.2021.3130902.

[14] E. Del Gobbo, A. Guarino, B. Cafarelli, and L. Grilli, "GradeAid: A framework for automatic short answers grading in educational contexts—design, implementation and evaluation," Springer, 2023. doi: 10.1007/s10115-023-01892-9.

[15] L. Zhang, Y. Huang, X. Yang, and S. Yu, "An automatic short-answer grading model for semi-open-ended questions," *Interactive Learning Environments*, Taylor & Francis, 2022. doi: 10.1080/10494820.2019.1648300.

[16] M. Messer, N. C. C. Brown, M. Kölling, and M. Shi, "Automated grading and feedback tools for programming education: A systematic review," *ACM Transactions on Computing Education*, 2024. doi: 10.1145/3636515.

[17] J. C. Paiva, J. P. Leal, and Á. Figueira, "Automated assessment in computer science education: A state-of-the-art review," *ACM Transactions on Computing Education*, 2022. doi: 10.1145/3513140.

[18] C. Grévisse, "LLM-based automatic short answer grading in undergraduate medical education," *BMC Medical Education*, Springer, 2024. doi: 10.1186/s12909-024-06026-5.

[19] M. Beseiso, O. A. Alzubi, and H. Rashaideh, "A novel automated essay scoring approach for reliable higher educational assessments," *Journal of Computing in Higher Education*, Springer, 2021. doi: 10.1007/s12528-021-09283-1.

[20] A. K. Bhowmick, A. Jagmohan, A. Vempaty, and P. Dey, "Automating question generation from educational text," Springer Nature, 2023. doi: 10.1007/978-3-031-47994-6_38.

[21] X. Dong, X. Zhang, Z. Li, Q. Hou, J. Xue, and X. Li, "A literature review of research on question generation in education," *PeerJ Computer Science*, 2025. doi: 10.7717/peerj-cs.3203.

[22] H. W. Awalurahman, R. F. Aji, and I. Budi, "Transformer and large language models for automatic multiple-choice question generation: A systematic literature review," *IEEE Access*, 2025. doi: 10.1109/ACCESS.2025.3590423.

[23] E. Bernasconi, D. Redavid, and S. Ferilli, "Enhancing personalised learning with a context-aware intelligent question-answering system and automated frequently asked question generation," *Electronics*, MDPI, 2025. doi: 10.3390/electronics14071481.

[24] Y. Song, J. Du, and Q. Zheng, “Automatic item generation for educational assessments: A systematic literature review,” *Interactive Learning Environments*, Taylor & Francis, 2025. doi: 10.1080/10494820.2025.2482588.

[25] E. Emekli and B. Karahan, “Artificial intelligence in radiology examinations: A psychometric comparison of question generation methods,” *Diagnostic and Interventional Radiology*, 2025. doi: 10.4274/dir.2025.253407.

[26] L. Ouahrani and D. Bennouar, “Paraphrase generation and supervised learning for improved automatic short answer grading,” *International Journal of Artificial Intelligence*, Springer, 2024. doi: 10.1007/s40593-023-00391-w.

[27] A. S. Abdullah, S. Geetha, A. B. A. Aziz, and U. Mishra, “Design of automated model for inspecting and evaluating handwritten answer scripts: A pedagogical approach with NLP and deep learning,” *Alexandria Engineering Journal*, Elsevier, 2024. doi: 10.1016/j.aej.2024.08.067.

[28] P. Sree Lakshmi, J. B. Simha, and R. Ranjan, “Empowering educators: Automated short answer grading with inconsistency check and feedback integration using machine learning,” *SN Computer Science*, Springer, 2024. doi: 10.1007/s42979-024-02954-7.

[29] M. Zachariou, O. Arandjelović, and D. J. Sloan, “Automated methods for tuberculosis detection/diagnosis: A literature review,” *BioMedInformatics*, MDPI, 2023. doi: 10.3390/biomedinformatics3030047.

[30] A. E. Kwabena, O. B. Wiafe, B. D. John, and A. Bernard, “An automated method for developing search strategies for systematic review using natural language processing (NLP),” *MethodsX*, Elsevier, 2023. doi: 10.1016/j.mex.2022.101935.